Particle identification using TMVA/MLP and Naïve Bayes for EMC detector

Malgorzata Gumberidze

IPN Orsay, France

Toolkit for Multivariate Data Analysis (TMVA)

- ✓ large variety of sophisticated data selection algorithms
  - ➢ Rectangular cut optimization
  - ➢ Projective and Multi-dimensional cut optimization
  - ➢ Fisher discriminant
  - ➢ ANN (3 diff. implementations)
  - ➢ Boosted/bagged Decision Trees
- ✓ have one common interface to different MVA method
  easy to use & to compare many different MVA methods
- ✓ common preprocessing of input data: decorrelation, PCA
- ✓ TMVA provides training/test and evaluation of all MVAs
- ✓ Each MVA method provides a ranking of input variables
- ✓ choose the best one for your selection problem–
- ✓ available as open source package
- ✓ however, still under development ... easily out of date

- *T*MVA is a sourceforge (SF) package for world-wide access
  - Home page .... http://tmva.sf.net/
  - SF project page .... http://sf.net/projects/tmva
  - View CVS .... http://tmva.cvs.sf.net/tmva/TMVA/
  - Mailing list ....http://sf.net/mail/?group_id=152074
  - Tutorial TWiki ....https://twiki.cern.ch/twiki/bin/view/TMVA/WebHome

MVA methods included in TMVA

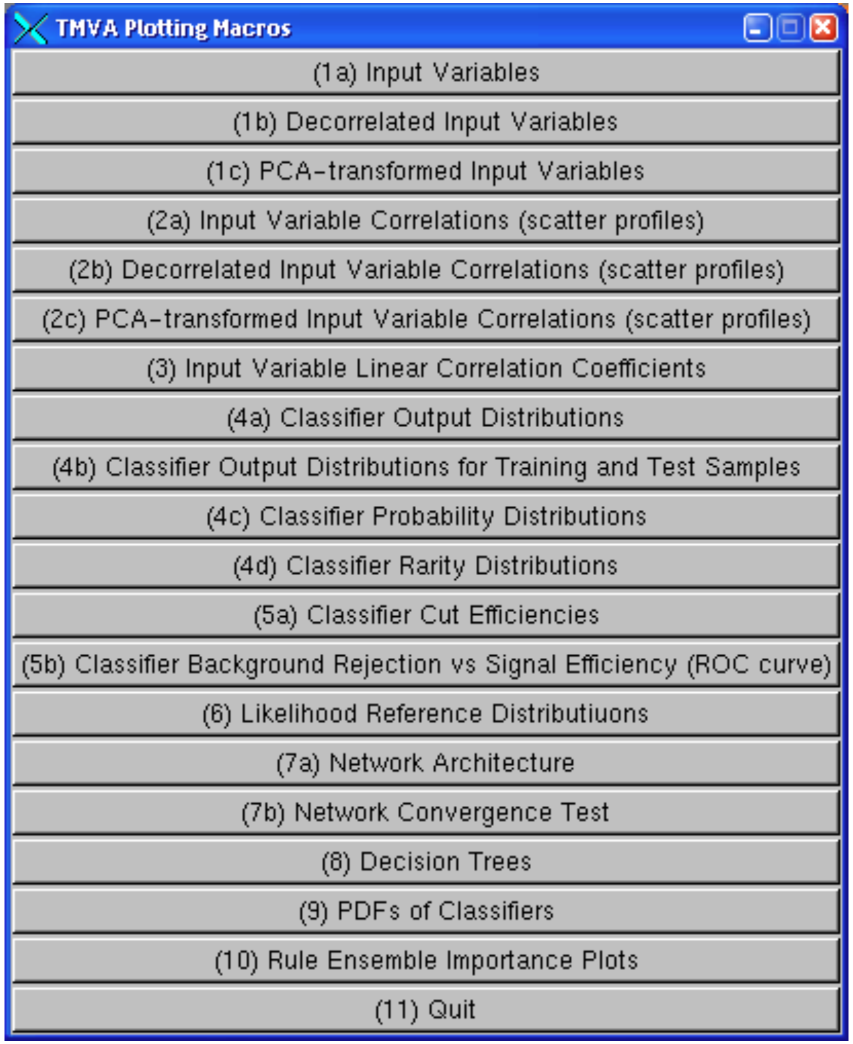
- Rectangular cut optimization
- Projection likelihood estimation ➔ studied independently by R. Kunne
- Multidimensional probability density estimation
  - Probability density estimator range search (PDERS)
  - Multidimensional K-Nearest Neighbour (K-NN) ➔ studied previously by M. Babai
- Linear discriminant analysis
  - H-Matrix (χ^2) Estimator
  - Fisher Discriminant
  - Function Discriminant Analysis (FDA)
- Boosted/Bagged decision trees (BDT)
- Artificial neural networks (ANN)
  - Clermont-Ferrand neural network
  - ROOT neural network
  - Multilayer Perceptron (MLP) neural network ➔ used previously in 'Babar framework'
- Predictive learning via rule ensemble (Rule-Fit)
- Support Vector Machine (SVM)

No single good classifier …

Criteria		Cuts	Likeli-hood	PDERS/ k-NN	H-Matrix	Fisher	MLP	BDT	RuleFit	SVM
		Classifiers								
Perfor-mance	no / linear correlations	😐	🙂	🙂	😐	🙂	🙂	😐	🙂	🙂
	nonlinear correlations	😐	☹️	🙂	☹️	☹️	🙂	🙂	😐	🙂
Speed	Training	☹️	🙂	🙂	🙂	🙂	😐	☹️	😐	☹️
	Response	🙂	🙂	☹️/😐	🙂	🙂	🙂	😐	😐	😐
Robust-ness	Overtraining	🙂	😐	😐	🙂	🙂	☹️	☹️	😐	😐
	Weak input variables	🙂	🙂	☹️	🙂	🙂	😐	😐	😐	😐
Curse of dimensionality		☹️	🙂	☹️	🙂	🙂	😐	🙂	😐	😐
Transparency		🙂	🙂	😐	🙂	🙂	☹️	☹️	☹️	☹️

TMVA evaluation tool

- TMVA is not only a collection of classifiers, but an MVA framework
- After training, TMVA provides ROOT evaluation scripts (through GUI)



Plot all signal (S) and background (B) input variables with and without pre-processing

Correlation scatters and linear coefficients for S & B

Classifier outputs (S & B) for test and training samples (spot overtraining)

Classifier *Rarity* distribution

Classifier significance with optimal cuts

B rejection versus S efficiency

Classifier-specific plots:
- Likelihood reference distributions
- Classifier PDFs (for probability output and Rarity)
- Network architecture, weights and convergence
- Rule Fitting analysis plots

- Visualise decision trees

How to chose out input for the training

Choose input variables sensibly:

- ✓ Do not include variables that are badly simulated
- ✓ Avoid variables with high correlations among themselves
  - ➢ drop all but one
- ✓ Some input variables have no discriminative power
  - ➢ drop them, reduce dimensionality
- ✓ Transform strongly peaked distributions into smoother ones, using log(), for instance
- ✓ Transform all variable in similar numerical range

Choose architecture sensibly:

- ➢ start with simple architecture, increase complexity gradually

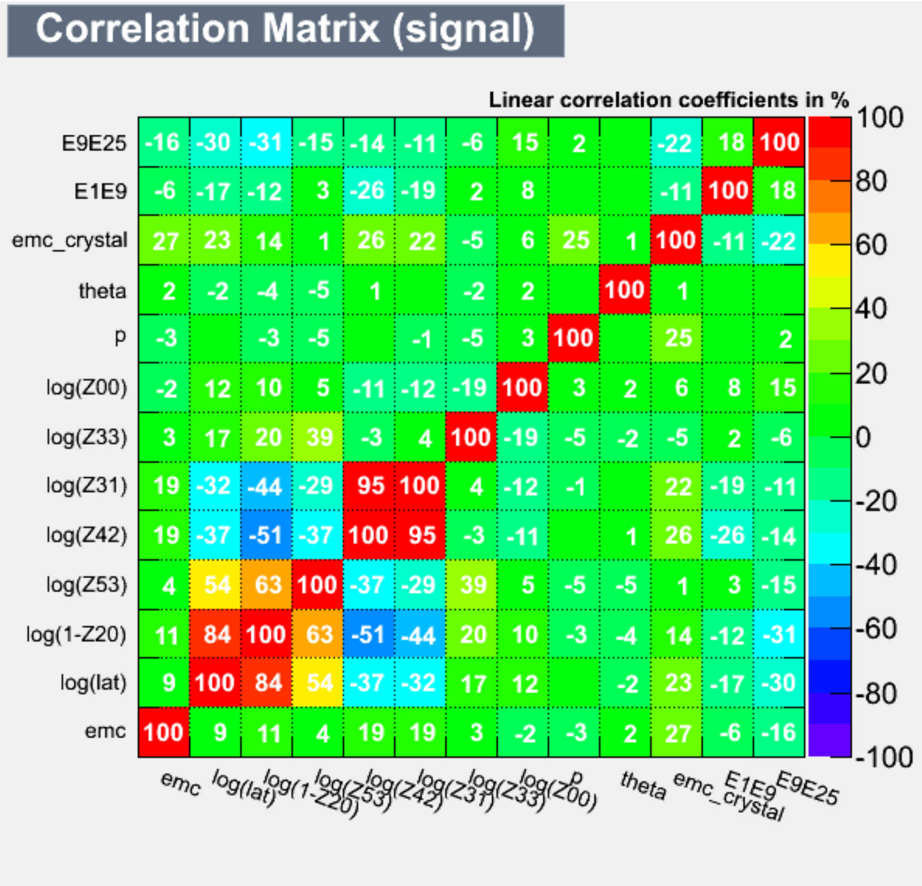
Avoid overtuning, use cross validation on independent training sample

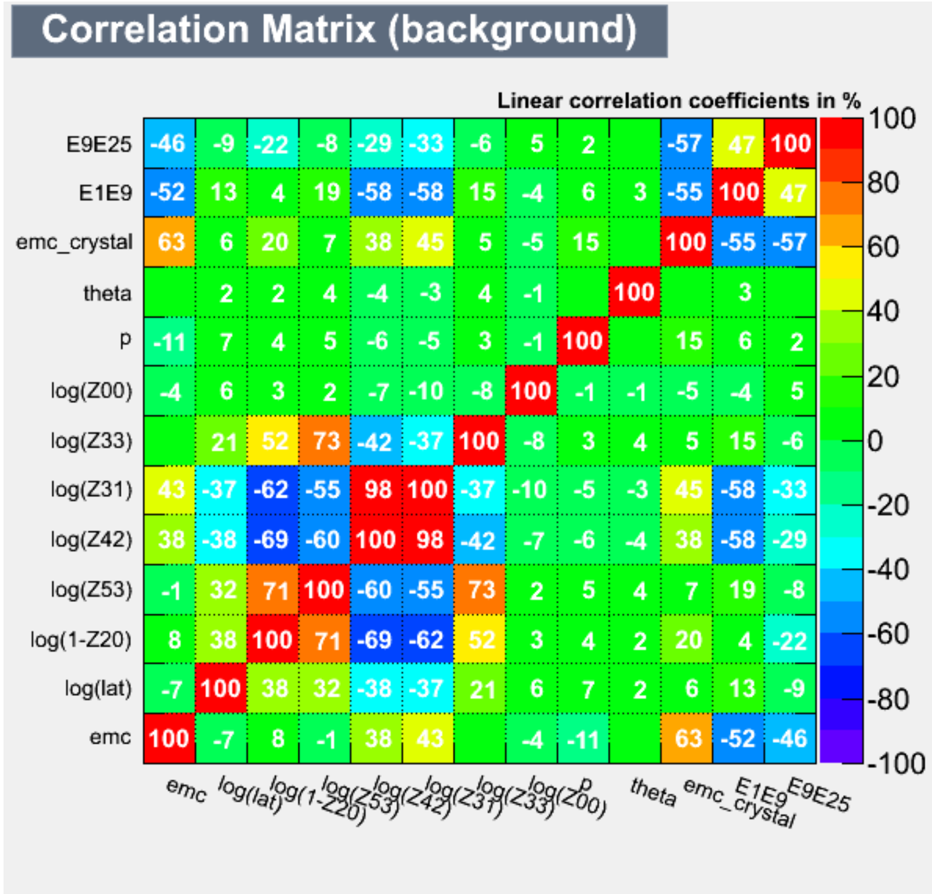
NN are no magic, understand what your trained NN is doing!

What is available from EMC detector …

electron

negative pion



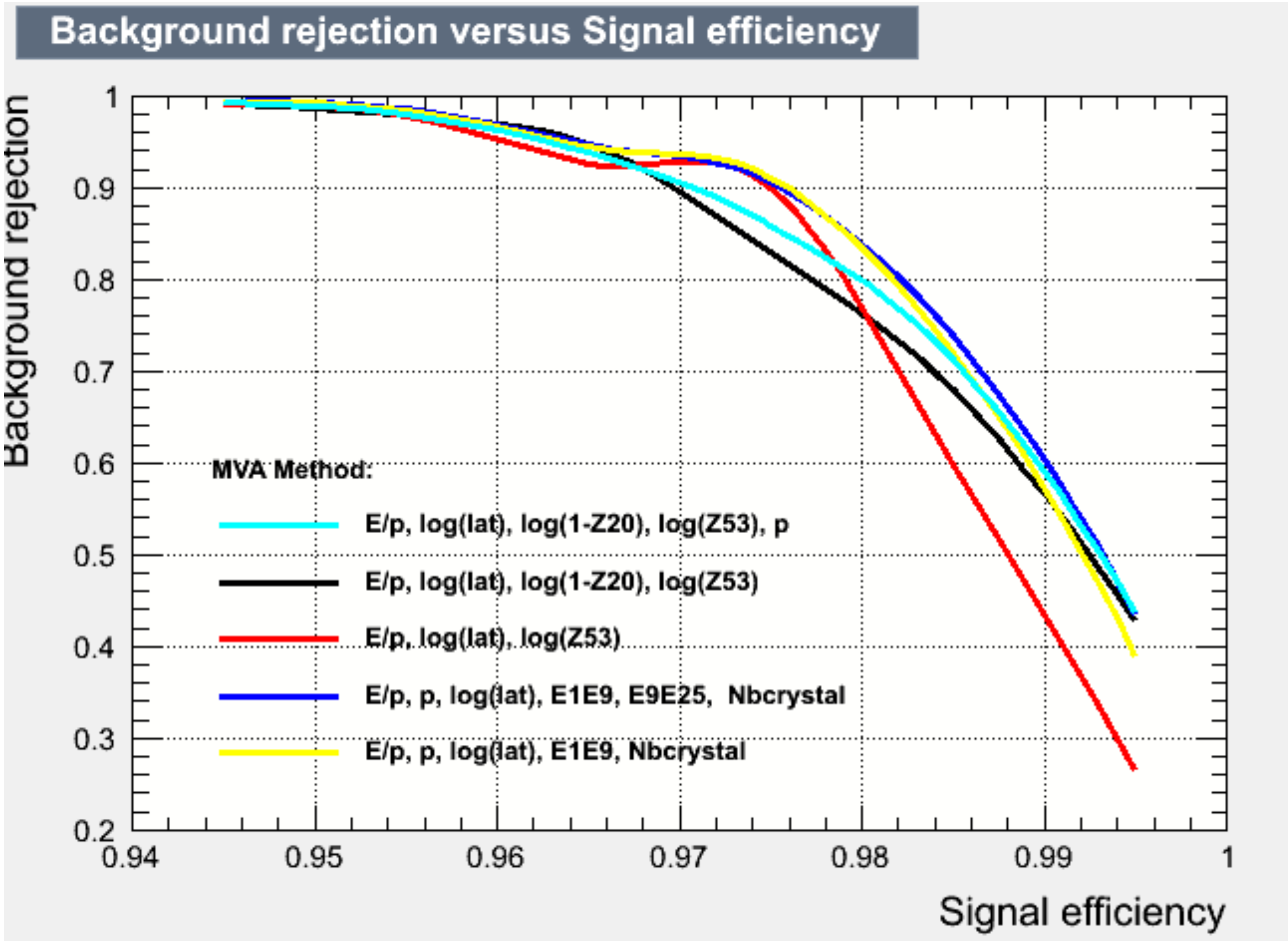


Monte Carlo momentum : 0.2 – 5 GeV/c

For the final PID following observables were selected:

E/p (emc), lateral momenta, E1/E9, E9/E25

ROC curve for different combination of parameters



Input, variables, conditions …

external: **january 2012**

pandaroot: **july 2012**

Testing done using 10^6 events : e^-, π^-, μ^-, K^-, p^-

Momentum range: 0.2 – 5 GeV/c

θ range: 5^o – 140^o

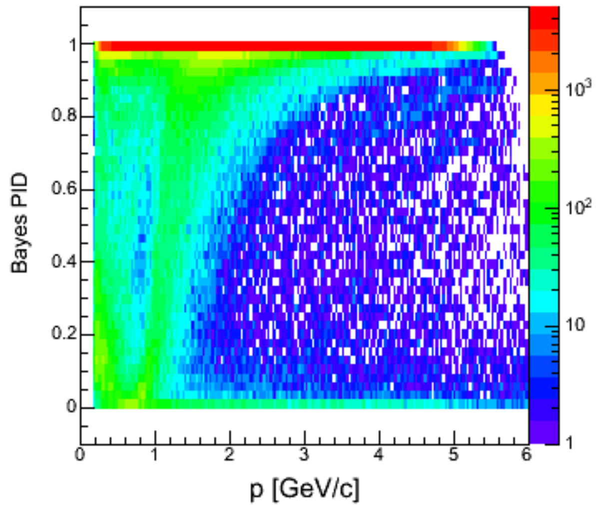
φ: 0^o – 360^o

MLP (MultiClass) trained on 10^5 events using**: Er/p, E1/E9, E9/E25, Lat**

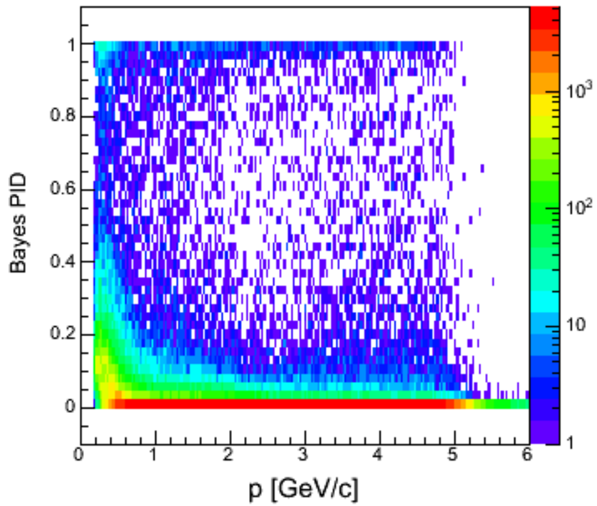
Naïve Bayes provided by Ronald **: Er/p, log(lat), log(Z53)**

PIDs from Naïve Bayes: momentum dependence

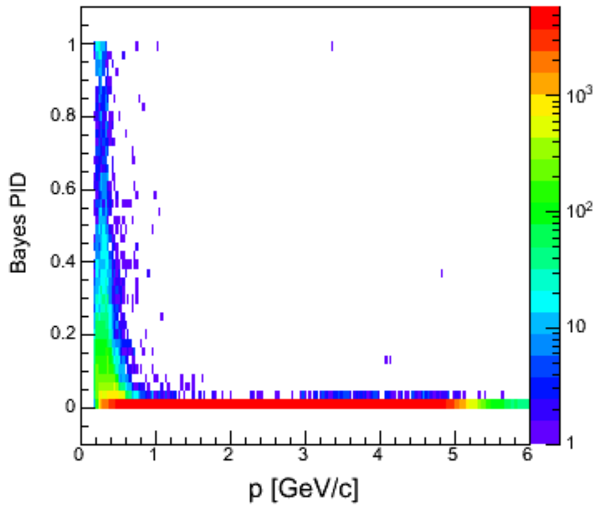
electron PID for electron



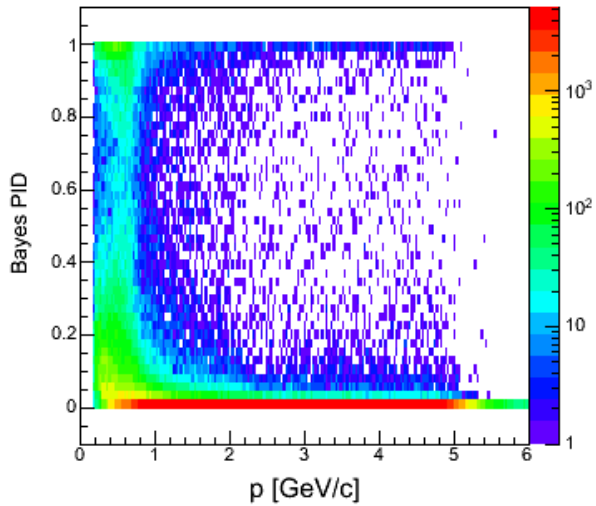
electron PID for π^-



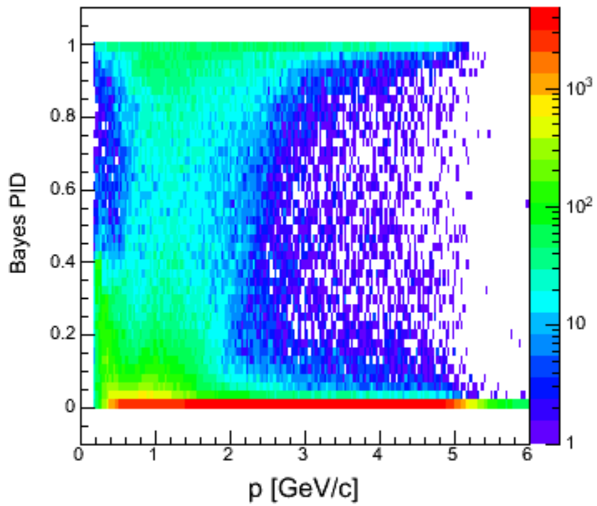
electron PID for μ^-



electron PID for K^-



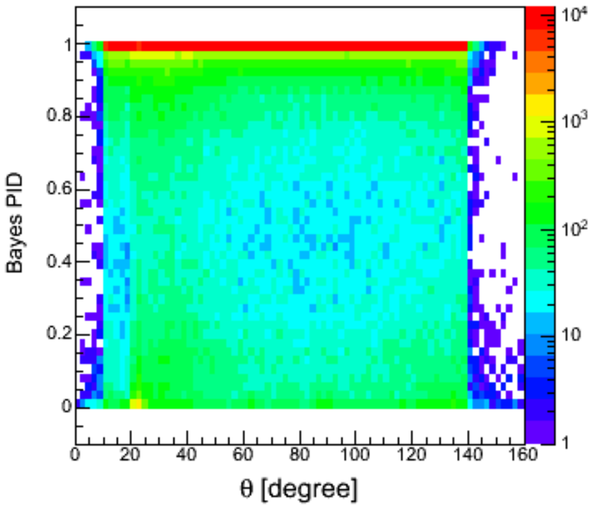
electron PID for p^-



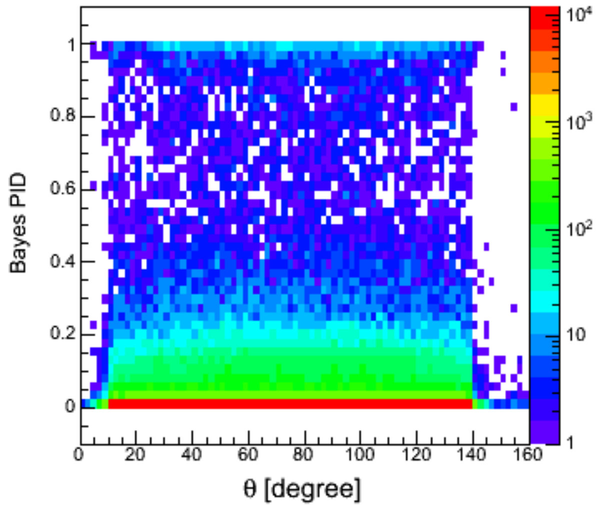
Shows momentum dependence: especially at low-momenta

PIDs from Naïve Bayes (II): θ dependence

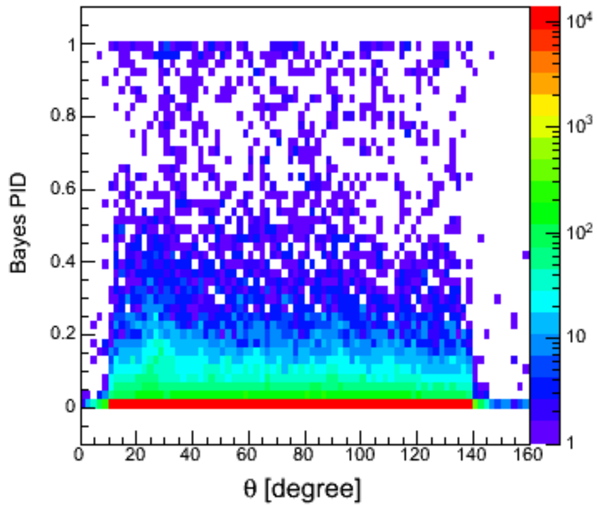
electron PID for electron



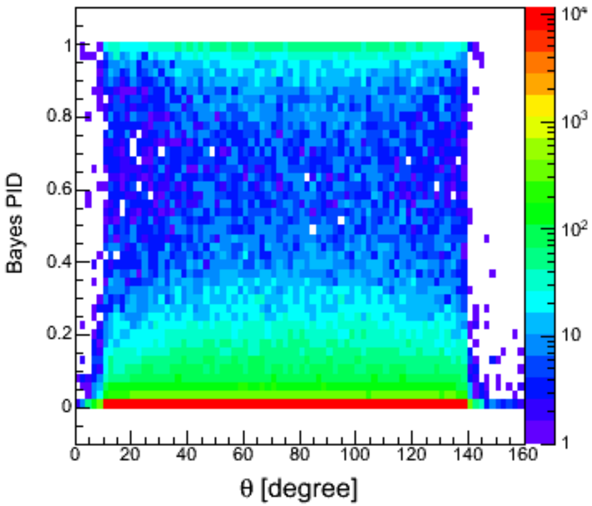
electron PID for π^-



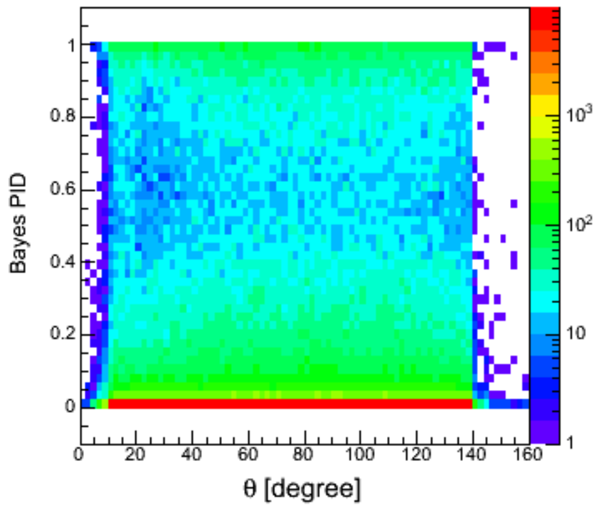
electron PID for μ^-



electron PID for K^-



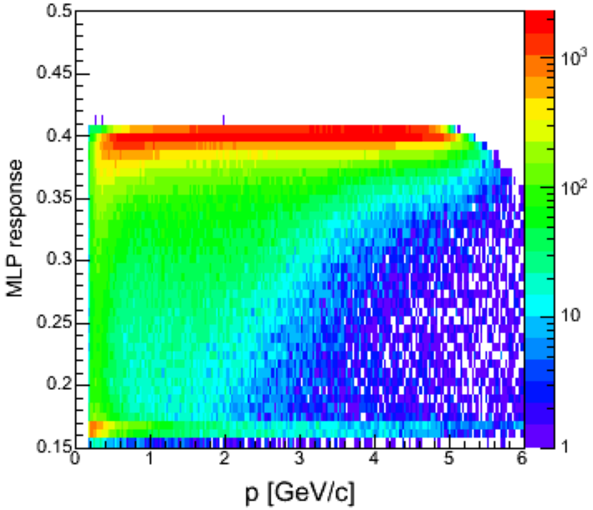
electron PID for p^-



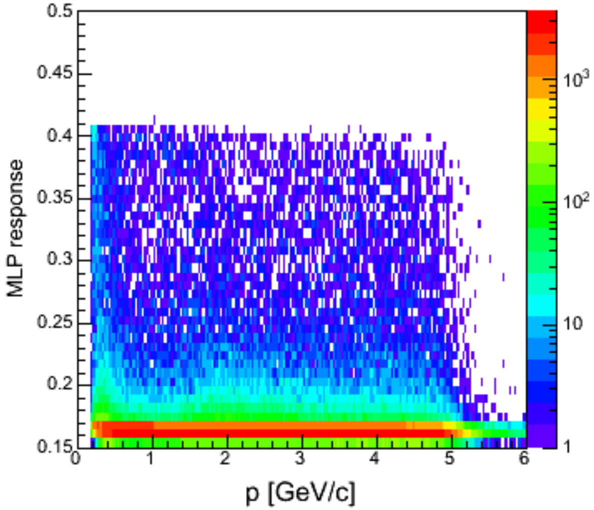
NO theta dependence

MLP response: momentum dependence

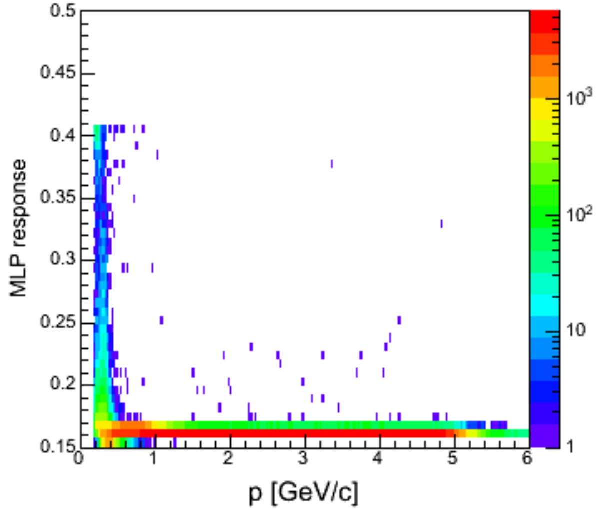
electron PID for electron



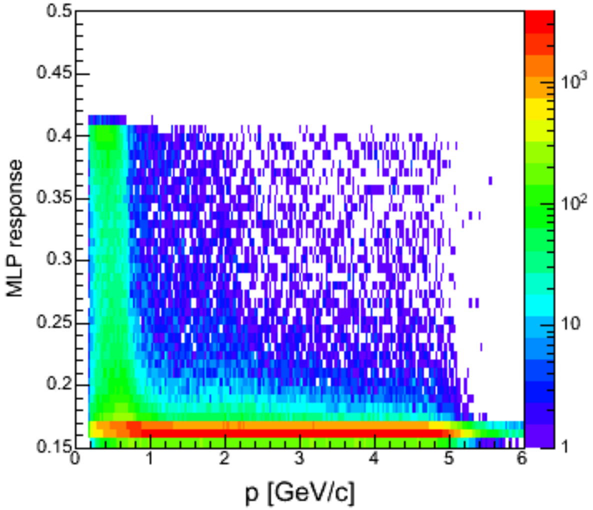
electron PID for π^-



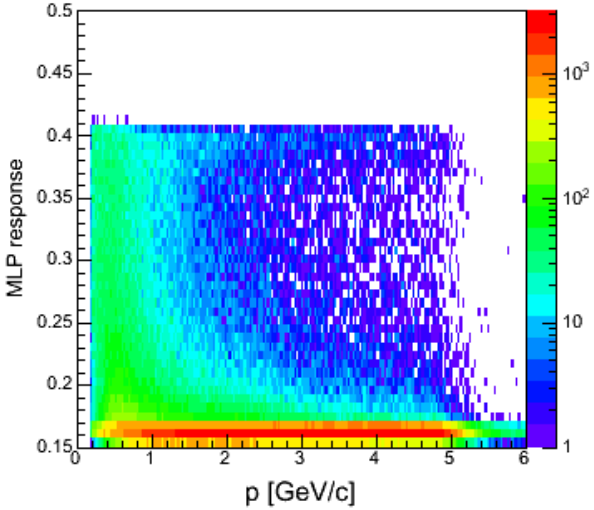
electron PID for μ^-



electron PID for K^-



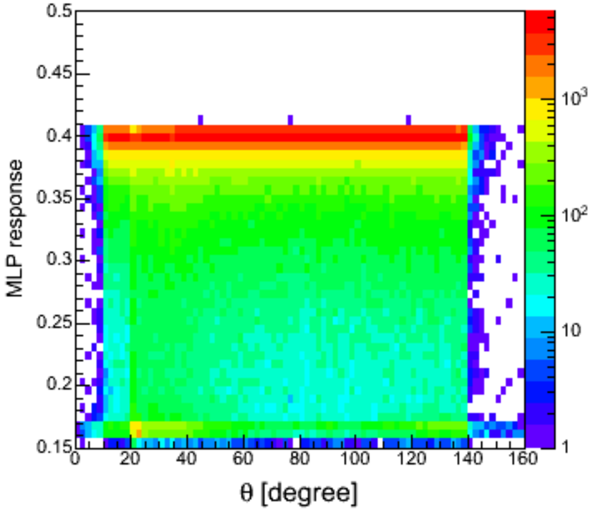
electron PID for p^-



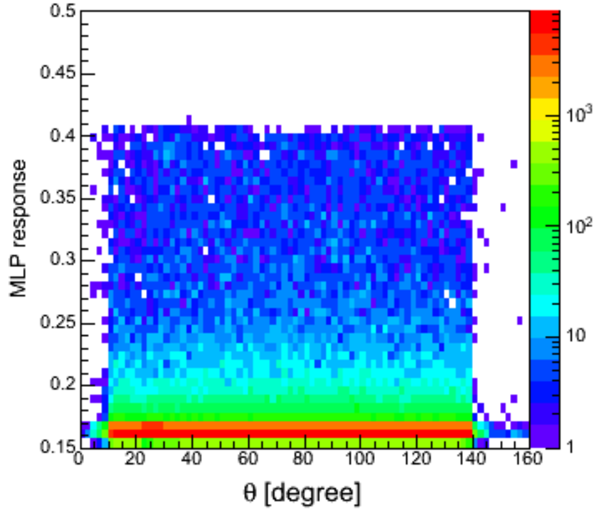
Similar momentum dependence as in bayes method

MLP response (II): momentum dependence

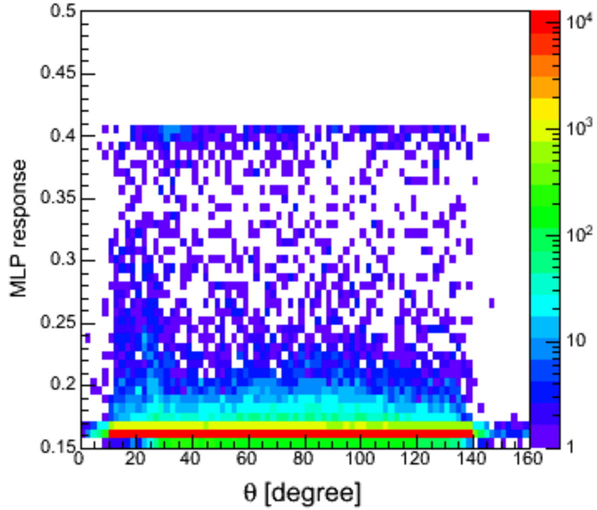
electron PID for electron



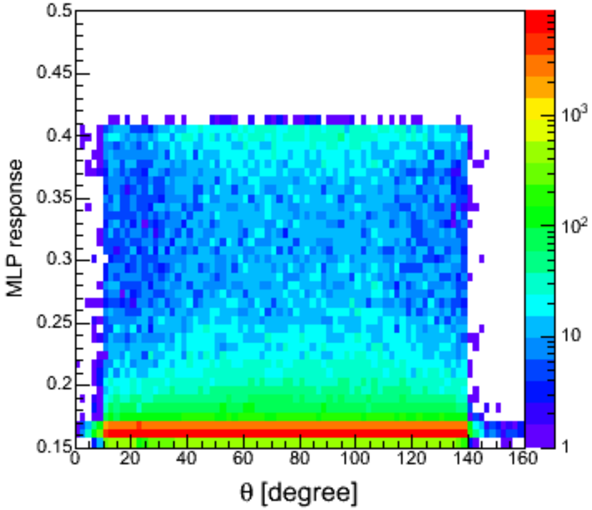
electron PID for π^-



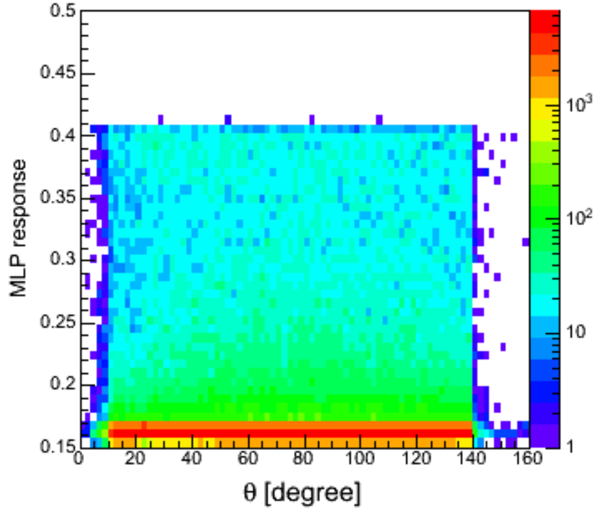
electron PID for μ^-



electron PID for K^-

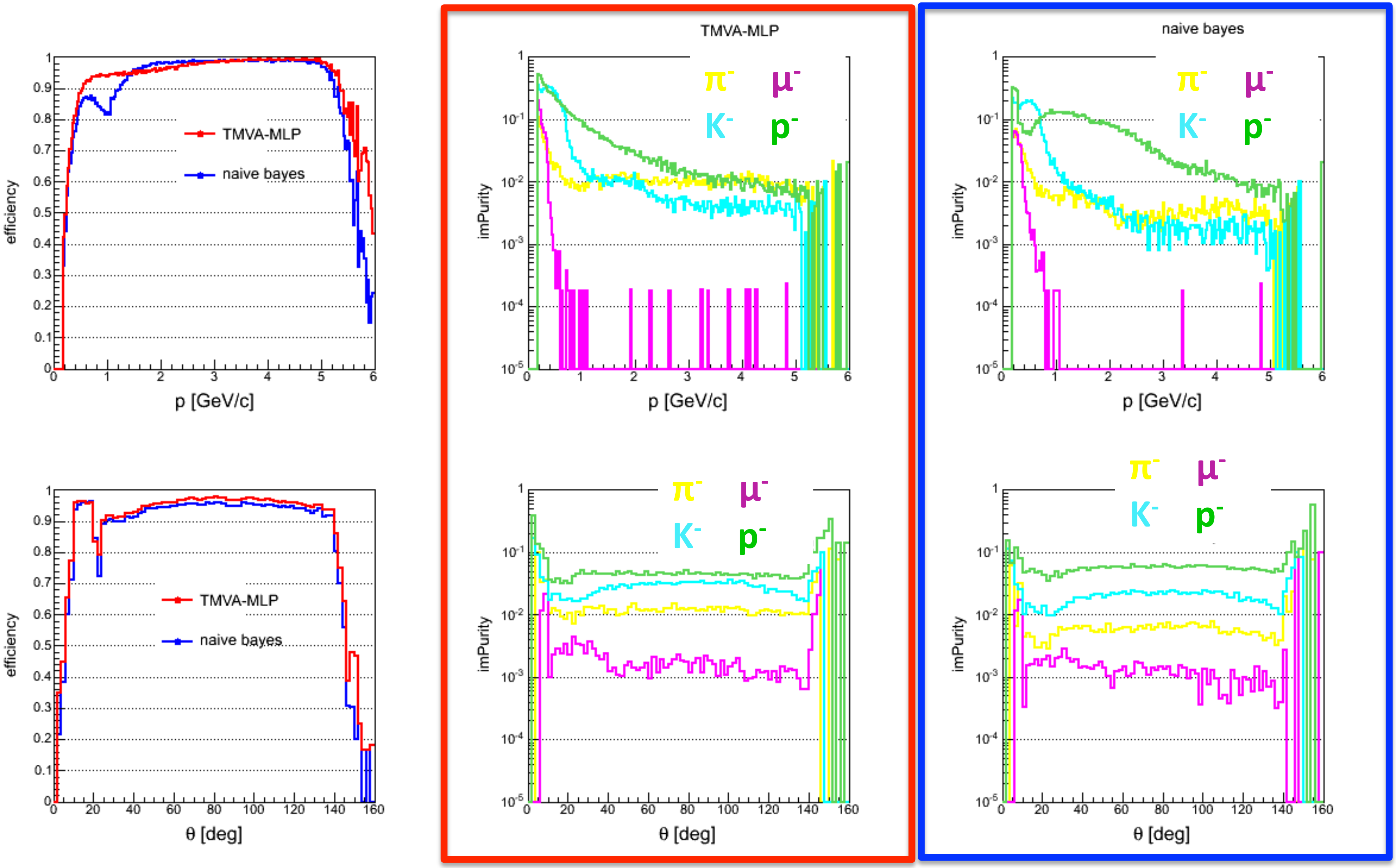


electron PID for p^-



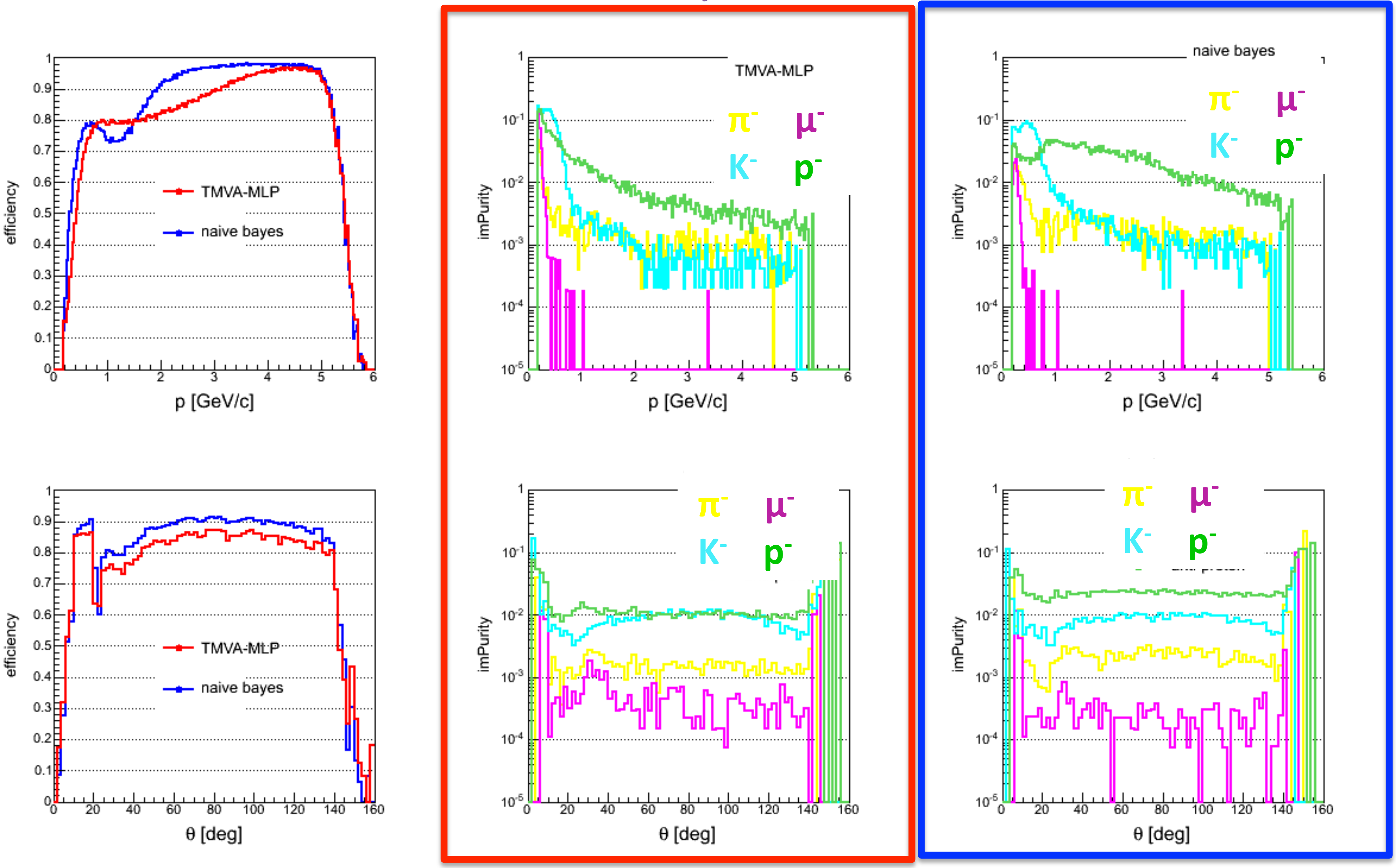
NO theta dependence

Comparison of the performance (I): best electron



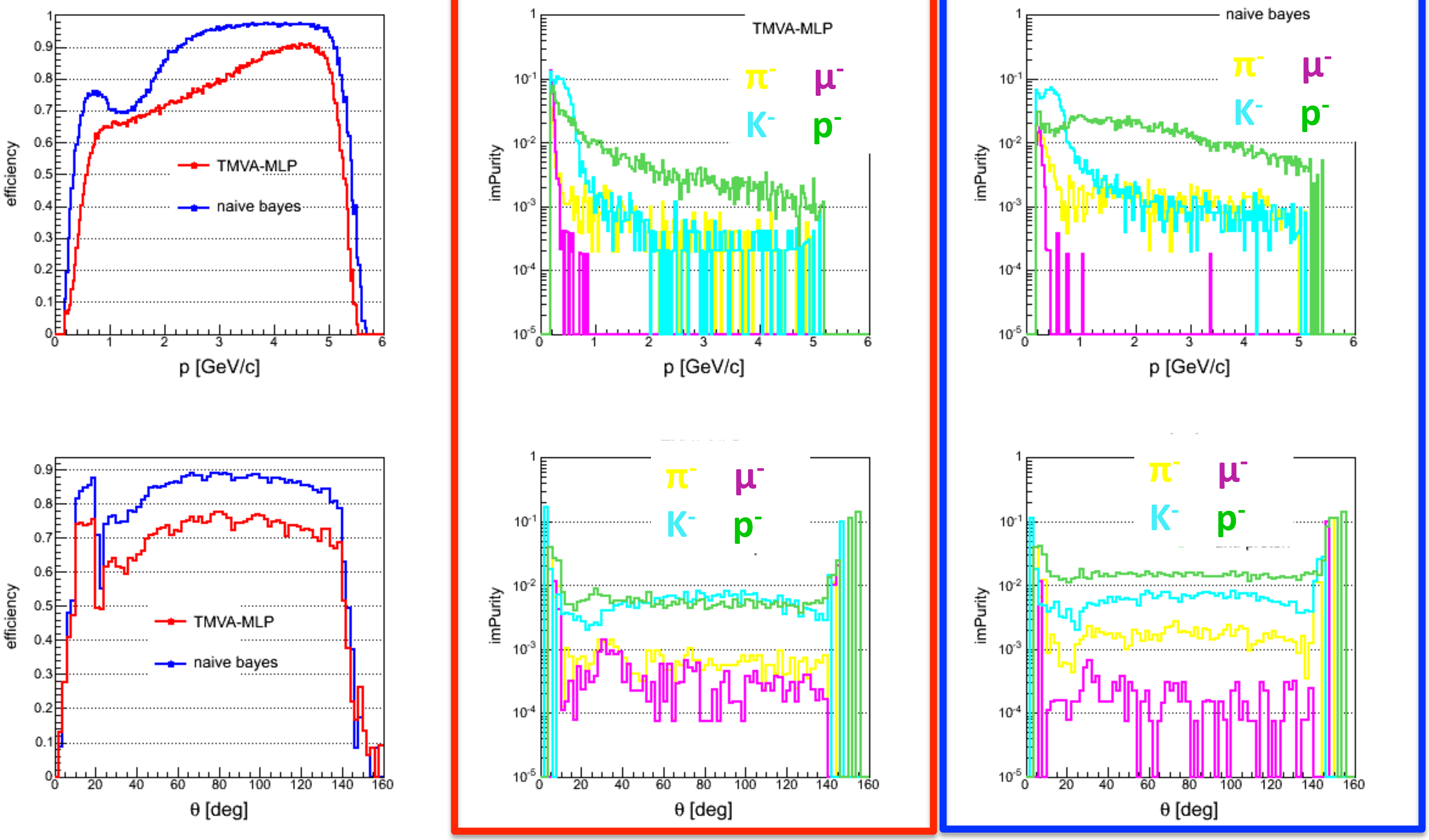
SELECTION: $(PID_e > PID_\pi) \&\& (PID_e > PID_\mu) \&\& (PID_e > PID_p) \&\& (PID_e > PID_k)$

Comparison of the performance (II): best electron, ELE > 90%



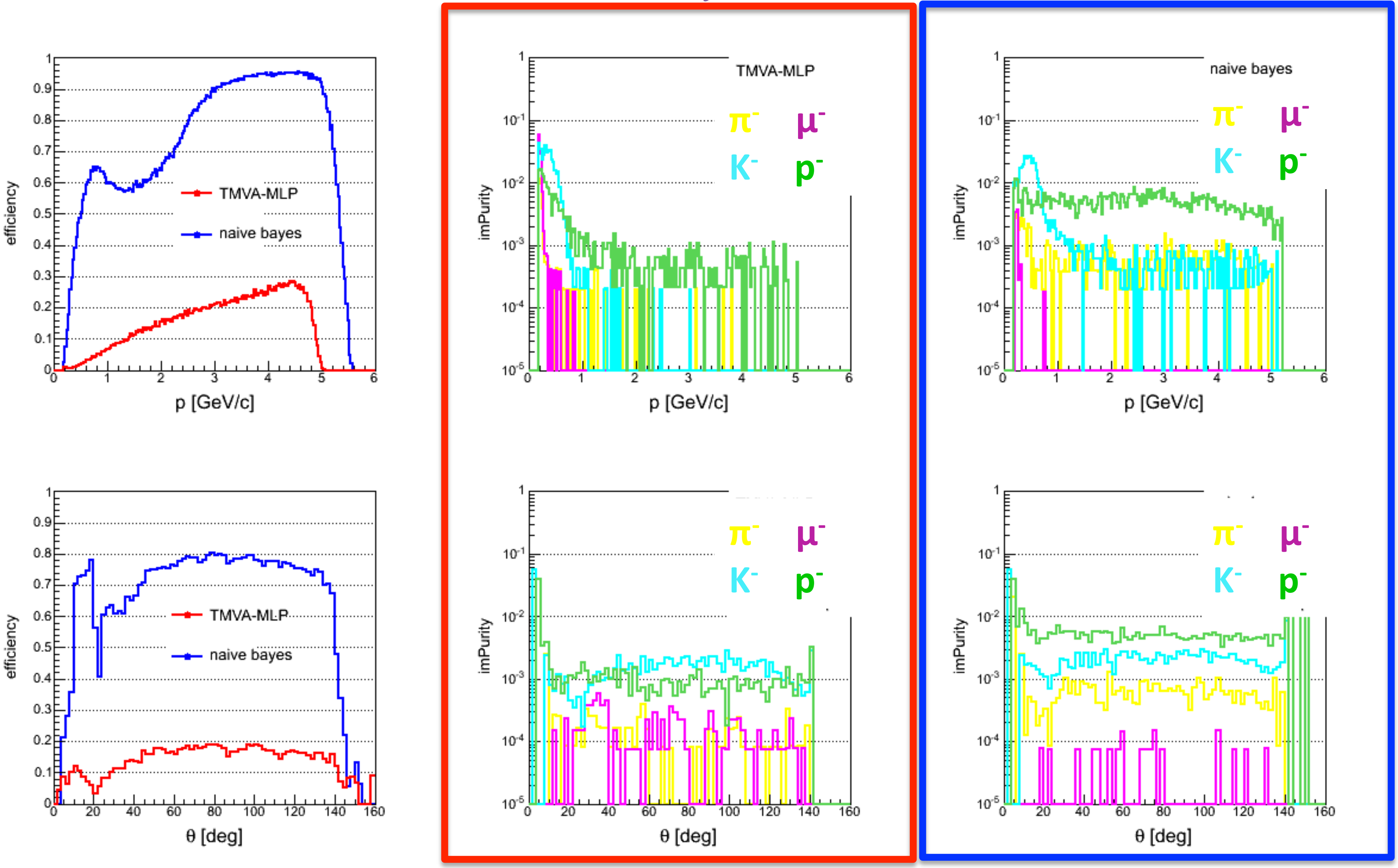
SELECTION: $(PID_e > PID_\pi)$ && $(PID_e > PID_\mu)$ && $(PID_e > PID_p)$ && $(PID_e > PID_k)$

Comparison of the performance (II): best electron, ELE > 95%



SELECTION: $(PID_e > PID_\pi)$ && $(PID_e > PID_\mu)$ && $(PID_e > PID_p)$ && $(PID_e > PID_k)$

Comparison of the performance (II): best electron, ELE > 99%



SELECTION: $(PID_e > PID_\pi) \&\& (PID_e > PID_\mu) \&\& (PID_e > PID_p) \&\& (PID_e > PID_k)$

Comparison of the PB and present analysis performance: ELE > 95%

Physics Book

has been calculated. Fig. 3.20 shows the electron efficiency and contamination rate as a function of momentum achieved by requiring an electron likelihood fraction of the EMC of more than 95 %. For momenta above 1 GeV/c one can see that the electron efficiency is greater than 98 % while the contamination by other particles is substantially less than 1 %. For momenta below 1 GeV/c, the electron

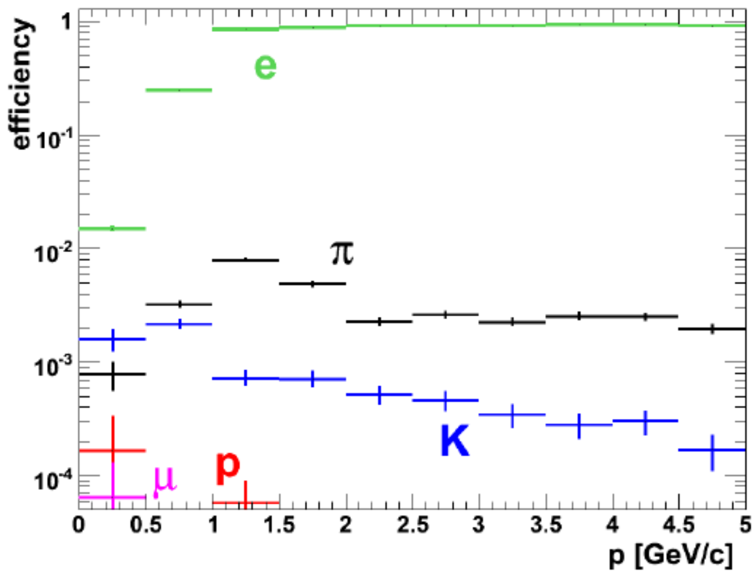


Figure 3.20: The electron efficiency and contamination rate for muons, pions, kaons and protons in different momentum ranges by using the EMC information.

10 input variables in total have been used, namely E/p, p, the polar angle θ of the cluster, and 7 shower shape parameters (E_1/E_9, E_9/E_{25}, the lateral moment of the shower and 4 Zernike moments). The

Comparison of the PB and present analysis performance inside PANDAroot: ELE > 95%

Physics Book

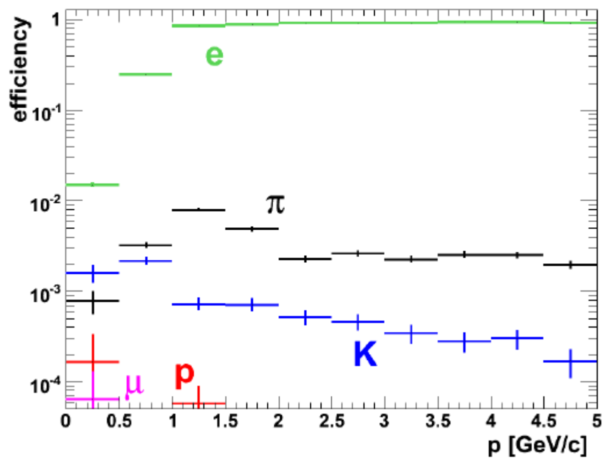


Figure 3.20: The electron efficiency and contamination rate for muons, pions, kaons and protons in different momentum ranges by using the EMC information.

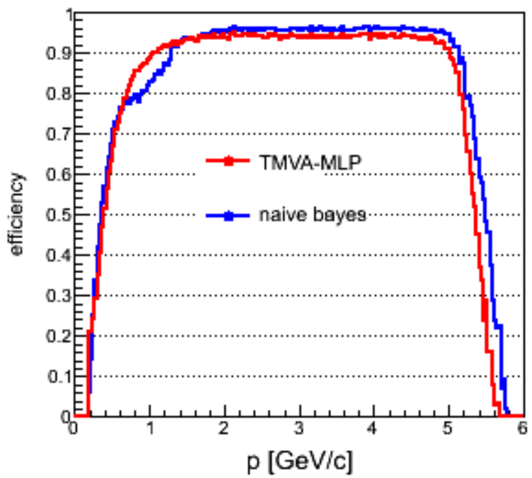
USING ONLY EMC information

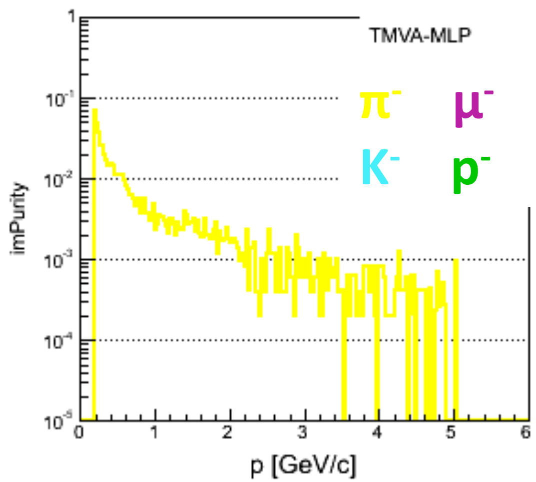
✓ **Electron efficiency:** using PandaRoot analysis methods (MLP and Bayes) we are able to reproduce Physics Book results

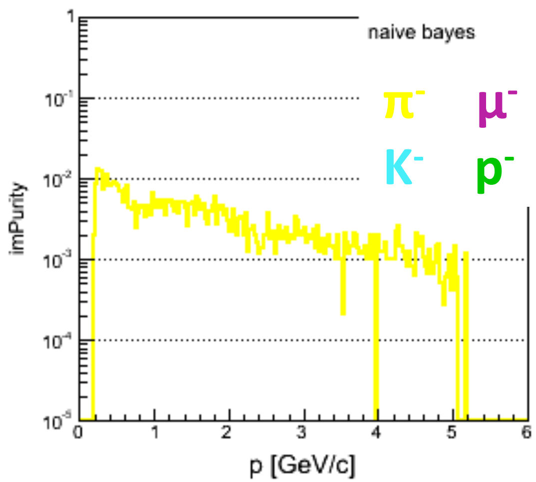
✓ **Pion impurity:**

✓ p > 2GeV models in PandaRoot shows smaller impurity

✓ for low momenta both PandaRoot models (MLP, Bayes) shows worst results. Still Bayes is better than MLP







Average efficiency and impurities

	MLP		naïve Bayes	
Signal efficiency				
Best e^-	95 %	(96%)	93 %	(93%)
Best e^- && 90%	82 %	(86%)	87 %	(87%)
Best e^- && 95%	71 %	(76%)	84 %	(84%)
Best e^- && 99%	15 %	(17%)	73 %	(73%)
Pion impurity				
Best e^-	1.16 %	(1.02%)	0.56 %	(0.56%)
Best e^- && 90%	0.15 %	(0.11%)	0.22 %	(0.21%)
Best e^- && 95%	0.066 %	(0.04%)	0.16 %	(0.15%)
Best e^- && 99%	0.01 %	(0.003%)	0.06 %	(0.06)

average over
full θ and
full p range

average over
full θ and
p > 0.7GeV/c

Summary and outlook

- ✓ Do we need to understand differences between BP and present MLP results ?
  - ✓ If yes, check if including the same variables as it was done for the Physics Book we also can obtain lower impurity for the EMC at lower momenta.
- ✓ Include information from other detectors: STT, DIRC, DISC into MLP -> on-going work
- ✓ Apply parameters (MLP) into the analysis of e^+e^- and $\pi^+\pi^-$